ehp ENVIRONMENTAL HEALTH PERSPECTIVES
http://www.ehponline.org

Instruments for Assessing Risk of Bias and Other Methodological Criteria of Published Animal Studies: A Systematic Review

David Krauth, Tracey J. Woodruff and Lisa Bero

http://dx.doi.org/10.1289/ehp.1206389

Online 14 June 2013

NIH National Institute of Environmental Health Sciences

Instruments for Assessing Risk of Bias and Other Methodological Criteria of Published Animal Studies: A Systematic Review

David Krauth[1]; Tracey J. Woodruff [2, 3]; Lisa Bero[1, 4]

[1] Department of Clinical Pharmacy, University of California, San Francisco, California, USA

[2] Department of Obstetrics, Gynecology, and Reproductive Sciences, University of California, San Francisco, California, USA

[3] Program on Reproductive Health and the Environment (PHRE), Oakland, California, USA

[4] Institute for Health Policy Studies, University of California, San Francisco, California, USA

Corresponding author:

Lisa Bero, PhD Professor

Department of Clinical Pharmacy Institute for Health Policy Studies University of California, San Francisco

Box 0613, 3333 California St, Suite 420

San Francisco, CA, USA 94118

Phone: (415) 476-1067

Fax: (415) 502-0792

berol@pharmacy.ucsf.edu

Key Words: animal experimentation; risk of bias; systematic review; toxicology; weight of evidence

Short Title: Risks of Bias in Animal Research

Acknowledgements

We acknowledge our funding source, the National Institute of Environmental Health Sciences (Grant # R21ES021028). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors also acknowledge Gloria Won (UCSF Mount Zion Campus) for her assistance with developing the search strategy. We also thank Dr. Dorie Apollonio (UCSF Laurel Heights Campus), Dr. David Dorman (North Carolina State University), and Rose Philipps (UCSF Laurel Heights Campus) for reviewing this manuscript.

The authors declare that they have no competing financial interests.

Abstract

Background: Results from animal toxicology studies are critical to evaluating potential harm from exposure to environmental chemicals or safety of drugs prior to human testing. However, there is significant debate about how to evaluate the methodology and potential biases of the animal studies. There is no agreed upon approach, and a systematic evaluation of current best practices is lacking.

Objective: We performed a systematic review to identify and evaluate instruments for assessing risk of bias and/or other methodological criteria of animal studies.

Method: We searched Medline (January 1966 - November 2011) to identify all relevant articles. We extracted data on risk of bias criteria (e.g. randomization, blinding, allocation concealment) and other study design features included in each assessment instrument.

Discussion: Thirty distinct instruments were identified with the total number of assessed risk of bias, methodological, and/or reporting criteria ranging from 2 to 25. The most common criteria assessed were randomization (25/30, 83%), investigator blinding (23/30, 77%) and sample size calculation (18/30, 60%). In general, authors failed to empirically justify why these or other criteria were included. Nearly all (28/30, 93%) instruments have not been rigorously tested for validity or reliability.

Conclusion: Our review highlights a number of risk of bias assessment criteria that have been empirically tested for animal research including randomization, concealment of allocation, blinding and accounting for all animals. Additionally, there is a need for empirically testing additional methodological criteria, and assessing the validity and reliability of a standard risk of bias assessment instrument.

Introduction

Results from animal toxicology studies are a critical, and often the only, input to evaluating potential harm from exposure to environmental chemicals or the safety of drugs before proceeding to human testing. However, there is significant debate about how to use animal studies in risk assessments and other regulatory decisions (Adami et al. 2011; ECETOC European Centre for Ecotoxicology and Toxicology of Chemicals 2009; Weed 2005; Woodruff and Sutton 2011). An important part of this debate is how to evaluate the methodology and potential biases of the animal studies in order to establish how confident one can be in the data.

For the evaluation of human clinical research, there is a distinction between assessing risk of bias and methodological quality (Higgins and Green 2008). Risks of bias are methodological criteria of a study that can introduce a systematic error in the magnitude or direction of the results (Higgins and Green 2008). In controlled human clinical trials testing the efficacy of drugs, studies with a high risk of bias, such as those lacking randomization, allocation concealment, or blinding of participants, personnel and outcome assessors produce larger treatment effect sizes, thus falsely inflating the efficacy of the drugs, compared to studies that have these design features (Schulz et al. 1995; Schulz and Grimes 2002a, b). Biased human studies assessing the harms of drugs are less likely to report statistically significant adverse effects (Nieto et al. 2007). An assessment of a study's methodology includes evaluation of additional study criteria related to how a study is conducted (e.g., in compliance with human subjects guidelines) or reported (e.g., study population described). Lastly, risk of bias is not the same as imprecision (Higgins and Green 2008). While bias refers to systematic error,

imprecision refers to random error. Although smaller studies are less precise, they may not be more biased.

While there is a well-developed and empirically-based literature on how to evaluate the risk of bias of randomized controlled clinical trials, less is known about how to do this for animal studies. Some risks of bias in animal studies have been identified empirically. For example, analyses of animal studies examining interventions for stroke, multiple sclerosis and emergency medicine have shown that lack of randomization, blinding, specification of inclusion/exclusion criteria, statistical power, and use of comorbid animals are associated with inflated effect estimates of pharmaceutical interventions (Bebarta et al. 2003; Crossley et al. 2008; Minnerup et al. 2010; Sena et al. 2010; Vesterinen et al. 2010). However, these studies have used a variety of instruments to evaluate the methodology of animal studies and often mix assessment of risks of bias, reporting, and other study criteria.

A number of guidelines and instruments for evaluating the risks of bias and other methodological criteria of animal research have been published, but there has been no attempt to compare the criteria that they include, to determine whether risk of bias, reporting or other criteria are assessed, or to determine if the criteria are based on empirical evidence of bias. The purpose of this paper is two-fold: (1) to systematically identify and summarize existing instruments for assessing risks of bias and other methodological criteria of animal studies and (2) to highlight the criteria that have been empirically tested for an association with bias in either animal or clinical models.

Methods

Inclusion/Exclusion Criteria

Articles that met the following inclusion criteria were included: (1) published report focusing on the development of an instrument for assessing the methodology of animal studies (2) English language. Where multiple analyses using a single instrument were published separately, the earliest publication was used. Modifications or updates of previously published instruments were considered new instruments and included. Applications of previously reported instruments, for example, to assess a certain area of animal research, were not included.

Search strategy

We searched Medline from January 1966 to November 2011 using a search term combination developed with input from expert librarians. Bibliographies from relevant articles were also screened to find any remaining articles that were not captured from the Medline search. Our search strategy contained the following MESH terms, text words and word variants:

((animal experimentation[mh] AND (standards[sh] OR research design[mh] OR bias[tw] OR biases[tw] OR checklist*[tw] OR translational research/ethics)) OR ((animals, laboratory[majr] OR disease models, animal[mh] OR drug evaluation, preclinical[mh] OR chemical evaluation OR chemical toxicity OR chemical safety) AND (research[majr:noexp] OR translational research[majr] OR research design[majr] OR "quality criteria") AND (guideline* OR bias[tw] OR biases[tiab] OR reporting[tw])) OR (animal*[ti] OR preclinical[ti] OR pre-clinical[ti] OR toxicology OR toxicological OR ecotoxicology OR

environmental toxicology AND (methodological quality OR research reporting OR study quality OR "risk of bias" OR "weight of evidence")) OR CAMARADES[tiab] OR "gold standard publication checklist" OR exclusion inclusion criteria animals bias) OR (peer review, research/standards AND Animals[Mesh:noexp])) OR (models, biological[mh] OR drug evaluation, preclinical[mh] OR toxicology[mh] OR disease models, animal[majr] AND (research design[mh] OR reproducibility of results[mh] OR "experimental design") AND (quality control[mh] OR guidelines as topic[mh] OR bias[tw] OR "critical appraisal") AND (Animals[Mesh:noexp])) AND eng[la]

Article selection

Studies were screened in two stages. Initially, abstracts and article titles were reviewed, and only those articles meeting our inclusion criteria were further scrutinized by reading the full text. Any articles that did not clearly meet the criteria after review of the full text were discussed by two authors and a decision was made about inclusion. Exact article duplicates were removed using Endnote X2 software.

Data extraction

We extracted data on each criterion included in each instrument and information on how the instrument was developed.

Instrument development and characteristics

We recorded the method used to develop each instrument (i.e. whether or not the criteria in the instrument were selected based on consensus, previous animal instruments, and/or clinical

instruments). We also recorded whether or not the criteria in the instrument were empirically tested to determine whether they were associated with biased effect estimates. Empirical testing was rated as done if at least one of the individual criterion was empirically tested.

Numerical methodological "quality" scores have been shown to be invalid for assessing risk of bias in clinical research (Juni et al. 1999). The current standard in evaluation of clinical research is to report each component of the assessment instrument separately and not calculate an overall numeric score (Higgins and Green 2008). Although the use of quality scores is now considered inappropriate, it is still a common practice. Therefore, we also assessed whether and how each instrument calculated a "quality" score.

We also noted whether or not the instrument was tested for reliability and validity. Reliability in assessing risk of bias refers to the extent to which results are consistent between different coders or in trials or measurements that are repeated (Carmines and Zeller 1979). Validity refers to whether the instrument measures what it was intended to measure, that is, methodological features that could affect research outcomes (Golafshani 2003).

Study design criteria to assess risk of bias and other methodological criteria

Based on published risk of bias assessment instruments for clinical research, we developed an *a priori* list of criteria and included additional criteria if they occurred in the review of the animal instruments (Cho and Bero 1994; Higgins and Green 2008; Jadad et al. 1996; Schulz et al. 2010).

We collected risk of bias, methodological, and reporting criteria as these three types of assessment criteria were often mixed in the individual instruments. The final list is shown below.

Treatment allocation/randomization. Describes whether or not treatment was randomly allocated to animal subjects so that each subject has an equal likelihood of receiving the intervention.

Concealment of Allocation. Describes whether or not procedures were used to protect against selection bias by ensuring that the treatment to be allocated is not known by the investigator before the subject enters the study.

Blinding. Relates to whether or not the investigator involved with performing the experiment, collecting data, and/or assessing the outcome of the experiment was unaware of which subjects received the treatment and which did not.

Inclusion/exclusion criteria. Describes the process used for including or excluding subjects.

Sample size calculation. Describes how the total number of animals used in the study was determined.

Compliance with animal welfare requirements. Describes whether or not the research investigators complied with animal welfare regulations.

Financial conflict of interest. Describes if the investigator(s) disclosed whether or not he/she has a financial conflict of interest.

Statistical model explained. Describes whether the statistical methods used and the unit of analysis are stated and whether the statistical methods are appropriate to address the research question.

Use of animals with comorbidity. Describes whether or not the animals used in the study have one or more pre-existing conditions that place them at greater risk of developing the health outcome of interest or responding differently to the intervention relative to animals without that condition.

Test animal descriptions. Describes the test animal characteristics including, the animal species, strain, sub-strain, genetic background, age, supplier, sex, and weight. At least one of these characteristics must be present for this criterion to be met.

Dose / response model. Describes whether or not an appropriate dose-response model was used given the research question and disease being modeled.

All animals accounted for. Describes whether or not the investigator accounts for attrition bias by detailing when animals were removed from the study and for what reason they were removed.

Optimal time window investigated. Describes whether or not the investigator provided sufficient time to pass before assessing the outcome. The optimal time window used in animal research should reflect the time needed to see the outcome and will depend on the hypothesis being tested. The optimal time window investigated should not be confused with the *therapeutic time window of treatment* which is defined as the time interval following exposure or onset of disease during which an intervention can still be effectively administered (Candelario-Jalil et al. 2005).

We extracted data on the study design criteria assessed by each instrument. We recorded the number of criteria assessed for each instrument, excluding criteria related only to journal reporting requirements (i.e., headers in an abstract).

Analysis

We report the frequency of each criterion assessed, as well as the frequency of any additional criteria that were included in the instruments.

Results

As shown in Figure 1, we identified 3731 potentially relevant articles. After screening of article titles and abstracts, 88 citations were identified for full text evaluation. After reviewing full text, 60 papers were excluded for at least one of three reasons: (1) not meeting inclusion criteria (2) studies reviewed preexisting instrument and (3) reported application of an instrument. After screening bibliographies, 2 additional instruments were found. Overall, 30 instruments were identified and included in the final analysis.

Table 1 lists the criteria of each instrument. Thirteen instruments were derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Three instruments were derived from previously developed clinically based risk of bias assessment instruments or citing clinical studies supporting the inclusion of specific criteria. Five instruments were developed using evidence from clinical research and either through consensus or citing past instrument publications. Three instruments were developed through consensus and citing past publications. Six instruments had no description of how they were developed.

Six instruments contained at least one criterion that showed an association of the criterion with inflated drug efficacy in animal models.

Seven instruments calculated a score for assessing methodological "quality". Descriptions of how these scores were calculated are provided in Table 1. Sixteen of the instruments were designed for no specific disease model and the most commonly modeled disease was stroke (9 out of 30 instruments).

Only one instrument was tested for validity (Sena et al. 2007) and one instrument was tested for reliability (Hobbs et al. 2005). Overall, 18 instruments were designed specifically to evaluate preclinical drug studies, 8 instruments documented general animal research guidelines, and 4 instruments were designed to assess environmental toxicology research.

The total number of risk of bias, methodological, and/or reporting criteria assessed by each instrument ranged from 2 to 25. Table 2 shows the study design criteria to assess risk of bias for each of the 30 instruments. Although these criteria were included in at least some of the instruments, they were not all supported by empirical evidence of bias. Blinding and randomization were the two most common criteria found in existing instruments; 25 instruments included randomization and 23 instruments included blinding. The need to provide a sample size calculation was listed in 18 instruments. None of the instruments contained all 13 criteria from our initial list; two instruments contained 9 criteria and four instruments contained only 1 or 2 of the criteria.

Additional criteria assessed by each instrument are listed in Table S1 (See Supplemental Material, Table S1). Some of these criteria related to reporting requirements for the abstract, introduction, methods, results and conclusions, rather than risk of bias criteria. These reporting criteria were not included in the count for the number of risk of bias criteria assessed by an

instrument. For example, Kilkenny et al. (2010) state that the ARRIVE Guidelines is a 20 criterion instrument. However, we listed the ARRIVE Guidelines as a 13 criterion instrument since 7 of the original criteria pertain to reporting requirements. Fourteen instruments contained criteria to describe animal housing, husbandry or physiological conditions. Inclusion of these criteria is empirically supported by studies showing that changes in housing conditions affect physiological and behavioral parameters in rodents (Duke et al. 2001; Gerdin et al. 2012). Among instruments that did not specify the need to use randomization, four out of five instruments still state that a control group should be used.

Discussion

This systematic review identified 30 instruments for assessing risk of bias and other methodological criteria of animal research. Identifying bias, the systematic error or deviation from the truth in actual results or inferences (Higgins and Green 2008), in animal research is important because animal studies are often the major or only evidence that forms the basis for regulatory or further research decisions. Our review highlights the variability in the development and content of instruments that are currently used to assess bias in animal research.

Most instruments were not tested for reliability or validity. One notable exception is the CAMARADES instrument developed by Sena and colleagues (Sena et al. 2007) that combined criteria from 4 previous instruments and showed that the instrument appears to have validity. Similarly, Hobbs et al. (2005) tested the reliability of a modified version of the Australasian ecotoxicity database (AED) based instrument and showed that it yielded an improvement in reliability relative to the original AED instrument. Furthermore, most of the instruments were

not developed based on empirical evidence showing an association between specific study design criteria and bias in research outcomes. Only 6 instruments included criteria that were supported by data showing an association between a particular methodological criterion and effect size in animal studies (Bebarta et al. 2003; Lucas et al. 2002; Macleod et al. 2004; Sena et al. 2007; Sniekers et al. 2008; Vesterinen et al. 2010). Most instruments contain criteria based on expert judgment. Others extrapolate from evidence of risk of bias in human studies. In addition, 7 instruments calculated a "quality score," although these scores are not considered a valid measure of risk of bias and this practice should be discontinued (Juni et al. 1999).

Biases that are known to influence the results of research include selection, performance, detection, and exclusion bias. These biases have been demonstrated in animal studies and methodological criteria that can protect against the biases have been empirically tested.

Selection bias, which introduces systematic differences between baseline characteristics in treatment and control groups, can be minimized by randomization and concealment of allocation. It has been shown empirically that lack of randomization or concealment of allocation in animal studies biases research outcomes by altering effect sizes (Bebarta et al. 2003; Macleod et al. 2008; Sena et al. 2007; Vesterinen et al. 2010). Performance bias is the systematic difference between treatment and control groups with regard to care or other exposure besides the intervention (Higgins and Green, 2008). Detection bias refers to systematic differences between treatment and control groups with regards to how outcomes are assessed (Higgins and Green, 2008). Blinding of investigators can protect against performance bias and there is substantial evidence that lack of blinding in a variety of types of animal studies is associated with exaggerated effect sizes (Bebarta et al. 2003; Sena et al. 2007; Vesterinen et al. 2010). Blinding

of outcome assessors is a primary way of reducing detection bias. There are many ways to achieve adequate blinding in animal studies, such as having coded data (thus, blinding to treatment assignment) analyzed by a statistician who is independent of the rest of the research team. Exclusion bias refers to the systematic difference between treatment and control groups in the number of animals that were included in and completed the study. Data on whether all animals in the study are accounted for and use of intention-to-treat analysis can reduce exclusion bias (Marshall et al. 2005).

Some criteria included in the animal research assessment instruments are not associated with bias. For example, a statement of compliance with animal welfare requirements is a reporting issue. Sample size calculations are often included as a criterion in animal research assessment instruments, but bias is not the same as imprecision. While bias refers to systematic error, imprecision refers to random error, meaning that multiple replications of the same study will produce different effect estimates because of sampling variation (Higgins and Green 2008). Although larger and more precise studies may give a more accurate estimate of an effect, they are not necessarily less biased. Furthermore, sample size calculations can be greatly affected by the underlying assumptions made for the calculation (Bacchetti 2010). Although a sample size calculation is not a risk of bias criterion, it is an important characteristic to consider when evaluating an overall body of evidence.

Some of the criteria listed in the instruments are unique to animal studies. For example, in preclinical drug research, testing animals with co-morbidities is necessary to identify whether or not candidate drugs retain efficacy in light of additional health complications and to more closely resemble the health status of humans. Empirical evidence supports the use of this criterion

because studies that included healthy animals instead of animals with comorbidities overestimated the effect sizes of experimental stroke interventions by over 10% (Crossley et al. 2008). For environmental chemicals, use of co-morbid animals could result in the opposite influence on effect size (i.e. to decrease it), and considering this as a criterion is consistent with recommendations to evaluate the influence of biological factors that may influence risk (EPA 2009). Timing of exposure also influences study outcome (Benatar 2007; van der Worp et al. 2010; Vesterinen et al. 2010), and some effects may only be observed for exposures that occur during certain developmental periods (EPA 2009). Gender, the nutritional status of experimental animals, and animal housing and husbandry conditions (Duke et al. 2001; Gerdin et al. 2012) could also impact the response to an intervention or environmental chemical exposure, but these criteria should be studied to determine if they introduce a systematic bias in results. These unique criteria have not been sufficiently included in the study instruments; and even if these criteria do not produce systematic bias, they should be clearly described and reported in animal studies to aid interpretation of the findings (Marshall et al. 2005).

Some risk of bias criteria have been investigated primarily in human studies, but warrant consideration for animal studies. Reviews of clinical studies have shown that study funding sources and financial ties of investigators (including university or industry affiliated investigators) are associated with favorable research outcomes for the sponsors (Lundh et al. 2011). Favorable research outcomes were defined as either increased effect sizes for drug efficacy studies, or decreased effect sizes for studies of drug harm. Selective reporting of outcomes and failure to publish entire studies is considered an important source of bias in clinical

studies, however, little is known about the extent of this bias in animal research (Hart et al. 2012; Rising et al. 2008).

Further research should consider potential interactions between risk of bias assessment criteria. Existing instruments have tested the association of study design criteria on effect size using univariate models. Multiple regression models should be performed to ascertain the relationship between a study design criterion and effect size when taking into account other criteria in the model. Covariance between methodological criteria should also be examined. For example, randomized studies may be less likely to omit blinding than non-randomized studies (van der Worp et al. 2010). Knowing the relative importance of these criteria will provide additional support for inclusion of specific criteria in risk of bias assessment instruments.

A majority of the instruments identified for our study exclude some criteria that appear to be important for assessing bias in animal studies (e.g. allocation concealment). It is important to recognize that some authors purposely exclude certain criteria from their instruments to reduce complexity and unnecessary detail. The most complex instrument had 25 criteria (Agerstrand et al. 2011). The detailed level of reporting needed to apply the Gold Standard Publication Checklist (GSPC), which has 17 criteria, was one of the main criticisms against it (Hooijmans et al. 2010).

As many journals now allow for online submission of supplemental data, risk of bias assessment should be less limited by a lack of space to report detailed methods. Reporting of clinical research has improved as risk of bias assessments for systematic reviews and other purposes became more prevalent and standards for reporting were implemented by journals (Turner et al.

2012). Recent calls for reporting criteria for animal studies (Landis et al. 2012; National Research Council (US) Institute for Laboratory Animal Research 2011) recognize the need for improved reporting of animal research. As happened for clinical research, reporting of animal research is likely to improve if risk of bias assessments become more common.

Many of the instruments identified in our review were derived to evaluate preclinical animal drug research which could limit their potential application in environmental health research. While selection, detection, and performance biases are relevant for all animal research, some of the preclinical instruments contain criteria specific for assessing the quality of stroke research, such as the "avoidance of anesthetics with intrinsic neuroprotective properties" (Macleod et al. 2004; Sena et al. 2007). On the other hand, investigation of an optimal time window for outcome assessment (EPA 2009), the timing of the exposure (EPA 2009), and measurement of outcomes that are sensitive to the exposure at the appropriate time (Wood 2000) are particularly important for assessing animal studies of environmental exposures.

Study limitations

A limitation of our study is that we may not have identified all published assessment instruments for animal research. As we limited our inclusion criteria to only articles published in English, it is possible that some published instruments may have been missed. Furthermore, we limited our search to articles indexed in Medline, so articles indexed exclusively in Embase or some other database would have been missed. However, our consultation with a librarian and the large pool of studies identified through the electronic search suggests that it was comprehensive.

Conclusions

Our review has identified a wide variety of instruments developed to evaluate animal studies. The individual criteria included in animal risk of bias assessment instruments should be empirically tested to determine their influence on research outcomes. Furthermore, testing these instruments for validity and reliability is needed. Lastly, there is a need to test existing instruments (many of which were developed using stroke models) on other animal models, to ensure their relevance and generalizability to other systems.

References

Adami HO, Berry SC, Breckenridge CB, Smith LL, Swenberg JA, Trichopoulos D, et al. 2011. Toxicology and epidemiology: improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. Toxicol Sci 122(2):223-234.

Agerstrand M, Kuster A, Bachmann J, Breitholtz M, Ebert I, Rechenberg B, et al. 2011. Reporting and evaluation criteria as means towards a transparent use of ecotoxicity data for environmental risk assessment of pharmaceuticals. Environ Pollut 159(10):2487-2492.

Altman DG, Gore SM, Garner MJ, Pocock SJ. 2000. Statistical guidelines for contributors to medical journals. 2nd Ed ed. London: BMJ Books.

Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. 2001. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 134(8):663-694.

Bacchetti P. 2010. Current sample size conventions: flaws, harms, and alternatives. BMC Med 8(17); doi:10.1186/1741-7015-8-17 (Online 22 March 2010).

Bebarta V, Luyten D, Heard K. 2003. Emergency medicine animal research: does use of randomization and blinding affect the results? Acad Emerg Med 10(6):684-687.

Benatar, M. 2007. Lost in translation: Treatment trials in the SOD1 mouse and in human ALS. Neurobiology of Disease 26(1):1-13.

Candelario-Jalil E, Mhadu NH, Gonzalez-Falcon A, Garcia-Cabrera M, Munoz E, Leon OS, et al. 2005. Effects of the cyclooxygenase-2 inhibitor nimesulide on cerebral infarction and neurological deficits induced by permanent middle cerebral artery occlusion in the rat. J Neuroinflammation 2(1):3.

Carmines EG, Zeller RA. 1979. Reliability and Validity Assessment: SAGE Publications, Inc.

Cho MK, Bero LA. 1994. Instruments for assessing the quality of drug studies published in the medical literature. JAMA 272(2):101-104.

Conrad JW, Jr., Becker RA. 2010. Enhancing credibility of chemical safety studies: emerging consensus on key assessment criteria. Environ Health Perspect 119(6):757-764.

Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, et al. 2008. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. Stroke 39(3):929-934.

de Aguilar-Nascimento JE. 2005. Fundamental steps in experimental design for animal studies. Acta Cir Bras 20(1):2-8.

Dirnagl U. 2006. Bench to bedside: the quest for quality in experimental stroke research. J Cereb Blood Flow Metab 26:1465–1478.

Duke JL, Zammit TG, Lawson DM. 2001. The effects of routine cage-changing on cardiovascular and behavioral parameters in male Sprague-Dawley rats. Contemp Top Lab Anim Sci 40(1):17-20.

Durda JL, Preziosi DV. 2000. Data Quality Evaluation of Toxicological Studies Used to

Derive Ecotoxicological Benchmarks. Human and Ecological Risk Assessment: An International Journal 6(5):747-765.

ECETOC (European Centre for Ecotoxicology and Toxicology of Chemicals B, Belgium). 2009. Framework for the Integration of Human and Animal Data in Chemical Risk Assessment. Technical Report No. 104 ISSN-0773-8072-104.

EPA Committee on Improving Risk Analysis Approaches Used by the U.S. EPA

National Research Council. 2009. Science and Decisions: Advancing Risk Assessment: The National Academies Press.

Festing MF. 2001. Guidelines for the design and statistical analysis of experiments in papers submitted to ATLA. Altern Lab Anim 29(4):427-446.

Festing MF. 2003. Principles: the need for better experimental design. Trends Pharmacol Sci 24(7):341-345.

Festing MF, Altman DG. 2002. Guidelines for the design and statistical analysis of experiments using laboratory animals. ILAR J 43(4):244-258.

Festing MF, van Zutphen LM. 1997. Guidelines for reviewing manuscripts on studies involving live animals. Synopsis of the workship. In: Animal alternatives, welfare and ethics. Amsterdam: Elsevier, 405-410.

Fisher M, Feuerstein G, Howells DW, Hurn PD, Kent TA, Savitz SI, et al. 2009. Update of the stroke therapy academic industry roundtable preclinical recommendations. Stroke 40(6):2244-2250.

Gerdin AK, Igosheva N, Roberson LA, Ismail O, Karp N, Sanderson M, et al. 2012. Experimental and husbandry procedures as potential modifiers of the results of phenotyping tests. Physiol Behav 106(5):602-611.

Golafshani N. 2003. Understanding Reliability and Validity in Qualitative Research. . The Qualitative Report 8(4):597-607.

Hart B, Lundh A, Bero L. 2012. Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses. BMJ 344:d7202.

Higgins JP, Green S. 2008. Cochrane Handbook for Systematic Reviews of Interventions. West Sussex, England: John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex.

Hobbs DA, Warne MSJ, Markich SJ. 2005. Evaluation of Criteria Used to Assess the Quality of Aquatic Toxicity Data. Integrated Environmental Assessment and Management 1(3):174 - 180.

Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. 2010. A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible. Altern Lab Anim 38(2):167-182.

Horn J, de Haan RJ, Vermeulen M, Luiten PG, Limburg M. 2001. Nimodipine in animal model experiments of focal cerebral ischemia: a systematic review. Stroke 32(10):2433-2438.

Hsu CY. 1993. Criteria for valid preclinical trials using animal stroke models. Stroke 24(5):633-636.

Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. 1996. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 17(1):1-12.

Johnson PD, Besselsen DG. 2002. Practical aspects of experimental design in animal research. ILAR J 43(4):202-206.

Jonas S, Ayigari V, Viera D, Waterman P. 1999. Neuroprotection against cerebral ischemia. A review of animal studies and correlation with human trial results. Ann N Y Acad Sci 890:2-3.

Juni P, Witschi A, Bloch R, Egger M. 1999. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 282(11):1054-1060.

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol 8(6):e1000412.

Klimisch HJ, Andreae M, Tillmann U. 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul Toxicol Pharmacol 25(1):1-5.

Lamontagne F, Briel M, Duffett M, Fox-Robichaud A, Cook DJ, Guyatt G, et al. 2010. Systematic review of reviews including animal studies addressing therapeutic interventions for sepsis. Crit Care Med 38(12):2401-2408.

Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. 2012. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490(7419):187-191.

Lucas C, Criens-Poublon LJ, Cockrell CT, de Haan RJ. 2002. Wound healing in cell studies and animal model experiments by Low Level Laser Therapy; were clinical studies justified? a systematic review. Lasers Med Sci 17(2):110-134.

Lundh A, Lexchin J, Sismondo S, Busuioc O, Bero L. 2011. Industry sponsorship and research outcome (Protocol). Cochrane Database of Systematic Reviews.

Macleod MR, Fisher M, O'Collins V, Sena ES, Dirnagl U, Bath PM, et al. 2009. Good laboratory practice: preventing introduction of bias at the bench. Stroke 40(3):e50-52.

Macleod MR, O'Collins T, Howells DW, Donnan GA. 2004. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke 35(5):1203-1208.

Macleod MR, van der Worp HB, Sena ES, Howells DW, Dirnagl U, Donnan GA. 2008. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. Stroke 39(10):2824-2829.

Markich SJ, Wame S, A-M W, C.J. R. 2002. A compilation of data on the toxicity of chemicals to species in Australia. Part 3: Metals. Australas J Exotoxicol 8:1-138.

Marshall JC, Deitch E, Moldawer LL, Opal S, Redl H, van der Poll T. 2005. Preclinical models of shock and sepsis: what can they tell us? Shock 24 Suppl 1:1-6.

Minnerup J, Heidrich J, Wellmann J, Rogalewski A, Schneider A, Schabitz WR. 2008. Meta-analysis of the efficacy of granulocyte-colony stimulating factor in animal models of focal cerebral ischemia. Stroke 39(6):1855-1861.

Minnerup J, Heidrich J, Rogalewski A, Schabitz WR, Wellmann J. 2009. The efficacy of erythropoietin and its analogues in animal stroke models: a meta-analysis. Stroke 40(9):3113-3120.

Minnerup J, Wersching H, Diederich K, Schilling M, Ringelstein EB, Wellmann J, et al. 2010. Methodological quality of preclinical stroke studies is not required for publication in high-impact journals. J Cereb Blood Flow Metab 30(9):1619-1624.

Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. 2005. Randomized trials stopped early for benefit: a systematic review. JAMA 294(17):2203-2209.

National Research Council (US) Institute for Laboratory Animal Research. 2011. Guidance for the Description of Animal Research in Scientific Publications 2012/03/02 ed. Washington D.C.: The National Academies Press.

Nieto A, Mazon A, Pamies R, Linana JJ, Lanuza A, Jimenez FO, et al. 2007. Adverse effects of inhaled corticosteroids in funded and nonfunded studies. Arch Intern Med 167(19):2047-2053.

Rice AS, Cimino-Brown D, Eisenach JC, Kontinen VK, Lacroix-Fralish ML, Machin I, et al. 2008. Animal models and the prediction of efficacy in clinical trials of analgesic drugs: a critical appraisal and call for uniform reporting standards. Pain 139(2):243-247.

Rising K, Bacchetti P, Bero L. 2008. Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation. PLoS Med 5(11):e217; discussion e217.

Schulz KF, Altman DG, Moher D. 2010. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 340:c332.

Schulz KF, Chalmers I, Hayes RJ, Altman DG. 1995. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273(5):408-412.

Schulz KF, Grimes DA. 2002a. Allocation concealment in randomised trials: defending against deciphering. Lancet 359(9306):614-618.

Schulz KF, Grimes DA. 2002b. Blinding in randomised trials: hiding who got what. Lancet 359(9307):696-700.

Sena ES, Briscoe CL, Howells DW, Donnan GA, Sandercock PA, Macleod MR. 2010. Factors affecting the apparent efficacy and safety of tissue plasminogen activator in thrombotic occlusion models of stroke: systematic review and meta-analysis. J Cereb Blood Flow Metab 30(12):1905-1913.

Sena E, van der Worp HB, Howells D, Macleod M. 2007. How can we improve the pre-clinical development of drugs for stroke? Trends Neurosci 30(9):433-439.

Simon R, Shiraishi K. 1990. N-methyl-D-aspartate antagonist reduces stroke size and regional glucose metabolism. Ann Neuorl 27(6):606-611.

Sniekers YH, Weinans H, Bierma-Zeinstra SM, van Leeuwen JP, van Osch GJ. 2008. Animal models for osteoarthritis: the effect of ovariectomy and estrogen treatment - a systematic approach. Osteoarthritis Cartilage 16(5):533-541.

STAIR. 1999. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. Stroke 30(12):2752-2758.

Turner L, Shamseer L., Altman DG, Schulz KF, Moher D. 2012. Does use of the CONSORT statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane Review. Systematic Reviews 1(60):1-7.

Unger EF. 2007. All is not well in the world of translational research. J Am Coll Cardiol 50(8):738-740.

van der Worp HB, de Haan P, Morrema E, Kalkman CJ. 2005. Methodological quality of animal studies on neuroprotection in focal cerebral ischaemia. J Neurol 252(9):1108-1114.

van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. 2010. Can animal models of disease reliably inform human studies? PLoS Med 7(3):e1000245.

Verhagen H, Aruoma OI, van Delft JH, Dragsted LO, Ferguson LR, Knasmuller S, et al. 2003. The 10 basic requirements for a scientific paper reporting antioxidant, antimutagenic or anticarcinogenic potential of test substances in in vitro experiments and animal studies in vivo. Food Chem Toxicol 41(5):603-610.

Vesterinen HM, Egan K, Deister A, Schlattmann P, Macleod MR, Dirnagl U. 2011. Systematic survey of the design, statistical analysis, and reporting of studies published in the 2008 volume of the Journal of Cerebral Blood Flow and Metabolism. J Cereb Blood Flow Metab 31(4):1064-1072.

Vesterinen HM, Sena ES, ffrench-Constant C, Williams A, Chandran S, Macleod MR. 2010 Improving the translational hit of experimental treatments in multiple sclerosis. Mult Scler 16(9):1044-1055.

Weed DL. 2005. Weight of evidence: a review of concept and methods. Risk Anal 25(6):1545-1557.

Wood, PA. 2000. Phenotype assessment: are you missing something? Comp Med. 50 (1):12-5.

Woodruff TJ, Sutton P. 2011. An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. Health Aff (Millwood) 30(5):931-937.

Table 1. Description of Instruments for Assessing Risk of Bias and Methodological Criteria of Animal Studies (n = 30)

Instrument Identifier	Method Used to Develop Instrument	Number of Criteria	Quality Score Calculated	Specific Disease Modeled	Instrument Criteria Empirically Tested	Intended Use of Instrument
Vesterinen et al. 2011	Method: Developed using evidence from clinical research and either through consensus or citing past animal instrument publications. Instrument development was based on previous research studies and new criteria not captured by past publications.	12	No	None	No	Preclinical drug research
Agerstrand et al. 2011	Method: Based on consensus and citing past guidelines. Authors collaborated with researchers and regulators to develop the criteria, relied on previously published reports, drew from their own professional experiences, and received additional suggestions from ecotoxicologists from Brixham Environmental Laboratories/AstraZeneca and researchers within the research programme MistraPharma.	25	No	None	No	Environmental toxicology research (specifically environmental risk assessment of pharmaceuticals)
National Research Council (US) Institute for Laboratory Animal Research 2011	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Evidence based rationale for including specific criteria is provided. Expert laboratory animal researchers with scientific publishing experience formed the committee that developed these guidelines.	19	No	None	No	General animal research

Instrument Identifier	Method Used to Develop Instrument	Number of Criteria	Quality Score Calculated	Specific Disease Modeled	Instrument Criteria Empirically Tested	Intended Use of Instrument
Lamontagne et al. 2010	Method: Developed using evidence from clinical research and either through consensus or citing past animal instrument publications. Relied on the PRISMA statement for determining relevant risk of bias criteria. Some of the criteria were incorporated into the risk of bias assessment based on clinical evidence showing an association between the criterion and overestimated treatment effect (Montori et al. 2005).	9	No	Sepsis	No	Preclinical drug research
Conrad and Becker 2010	Method: Consensus and citing past guidelines. Constructed using 5 previously developed quality assessment guidelines.	10	Yes[a]	None	No	General animal research
Vesterinen et al. 2010	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Derived from the consensus statement "Good laboratory Practice" for modeling stroke (Macleod et al. 2009).	5	No	Multiple Sclerosis	Yes	Preclinical drug research
Kilkenny et al. 2010 The ARRIVE Guidelines	Method: Developed using evidence from clinical research and either through consensus or citing past animal instrument publications. Developed using the CONSORT criteria, consensus and consultation among scientists, statisticians, journal editors and research funders.	13	No	None	No	General animal research
Minnerup et al. 2010	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Derived from the STAIR recommendations (STAIR 1999).	11	Yes[b]	Stroke	No	Preclinical drug research

Instrument Identifier	Method Used to Develop Instrument	Number of Criteria	Quality Score Calculated	Specific Disease Modeled	Instrument Criteria Empirically Tested	Intended Use of Instrument
Hooijmans et al. 2010 The Gold Standard Publication Checklist (GSPC)	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Many of the criteria in the GSPC are supported by previous studies showing the importance of such parameters. The authors also discussed and optimized the GSPC with animal science experts.	17	No	None	No	General animal research
van der Worp et al. 2010	Method: Developed using evidence from clinical research and either through consensus or citing past animal instrument publications. Recommendations based largely on CONSORT and to a smaller extent on animal guidelines (Altman et al. 2001; Dirnagl 2006; Macleod et al. 2009; Sena et al. 2007; STAIR 1999).	9	No	Stroke	No	Preclinical drug research
Macleod et al. 2009	Method: Developed using evidence from clinical research and either through consensus or citing past animal instrument publications. Criteria based on past meta-analyses done by CAMARADES researchers and CONSORT.	9	No	Stroke	No	Preclinical drug research
Fisher et al. 2009	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Updated the original (STAIR 1999) guidelines. No description of how the new instrument was developed.	15	No	Stroke	No	Preclinical drug research
Rice et al. 2008	Method: Derived from previously developed clinically based risk of bias assessment instruments or citing clinical studies supporting the inclusion of specific criteria. Modified form of the Jadad criteria used to assess clinical interventions.	6	No	Animal Pain Models	No	Preclinical drug research

Instrument Identifier	Method Used to Develop Instrument	Number of Criteria	Quality Score Calculated	Specific Disease Modeled	Instrument Criteria Empirically Tested	Intended Use of Instrument
Sniekers et al. 2008	No description of how the instrument was developed.	7	No	Osteoarthritis	Yes	Preclinical drug research
Sena et al. 2007	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Derived from 4 previous checklists: Stroke Therapy Academic Industry Roundtable (STAIR 1999); Amsterdam criteria (Horn et al. 2001); Collaborative Approach to Meta-Analysis and Review of Animal Data in Experimental Stroke (CAMARADES) (Macleod et al. 2004); and Utrecht criteria (van der Worp et al. 2005).	21	No	Stroke	Yes	Preclinical drug research
Unger 2007	No description of how the instrument was developed.	4	No	None	No	Preclinical drug research
Hobbs et al. 2005	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Modified version of Australasian ecotoxicity database (AED) quality assessment scheme (Markich et al. 2002).	18	Yes[c]	None	No	Environmental toxicology research
Marshall et al. 2005	Method: Derived from previously developed clinically based risk of bias assessment instruments or citing clinical studies supporting the inclusion of specific criteria. This instrument was based on CONSORT.	10	No	Shock/Sepsis	No	Preclinical drug research

Instrument Identifier	Method Used to Develop Instrument	Number of Criteria	Quality Score Calculated	Specific Disease Modeled	Instrument Criteria Empirically Tested	Intended Use of Instrument
van der Worp et al. 2005 Utrecht Criteria	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. The checklist was derived from the STAIR criteria (STAIR 1999) and recommendations also resemble the scale used by (Horn et al. 2001).	9	Yes	Stroke	No	Preclinical drug research
de Aguilar-Nascimento 2005	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Motivated by past research describing the importance of certain study design features (Festing 2003; Festing and Altman 2002; Johnson and Besselsen 2002).	9	No	None	No	General animal research
Macleod et al. 2004	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Informed by previous published criteria (Horn et al. 2001; Jonas et al. 1999).	10	Yes[d]	Stroke	Yes	Preclinical drug research
Bebarta et al. 2003	Method: Derived from previously developed clinically based risk of bias assessment instruments or citing clinical studies supporting the inclusion of specific criteria. Randomization and blinding were included based on evidence from human clinical trials showing that lack of these features often overestimates the magnitude of treatment effects.	2	No	None	Yes	Preclinical drug research
Verhagen et al. 2003	No description of how the instrument was developed.	10	No	None	No	General animal research

Instrument Identifier	Method Used to Develop Instrument	Number of Criteria	Quality Score Calculated	Specific Disease Modeled	Instrument Criteria Empirically Tested	Intended Use of Instrument
Festing and Altman 2002	Method: Developed based on consensus and citing past guidelines. Derived from published guidelines for contributors to medical journals (Altman et al. 2000), in-vitro models (Festing 2001), and a previously published checklist (Festing and van Zutphen 1997).	10	No	None	No	General animal research
Johnson and Besselsen 2002	No description of how the instrument was developed.	7	No	None	No	General animal research
Lucas et al. 2002	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. An 8-point rating system was developed based on 2 previous recommendations (Horn et al. 2001; STAIR 1999).	8	Yes[e]	None	Yes	Preclinical drug research
Horn et al. 2001 Amsterdam Criteria	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Derived partially from the original STAIR guidelines (STAIR 1999).	8	Yes[f]	Stroke	No	Preclinical drug research
Durda and Preziosi 2000	Method: Derived by modifying or updating previously developed animal research methodology assessment instruments or citing animal studies supporting the inclusion of specific criteria. Compiled methodological requirements and acceptance criteria for ecotoxicology testing published by national and international governmental and testing organizations.	15	No	None	No	Environmental toxicology research

Instrument Identifier	Method Used to Develop Instrument	Number of Criteria	Quality Score Calculated	Specific Disease Modeled	Instrument Criteria Empirically Tested	Intended Use of Instrument
Klimisch et al. 1997	No description of how the instrument was developed.	9	No	None	No	Environmental toxicology research
Hsu 1993	No description of how the instrument was developed.	6	No	Stroke	No	Preclinical drug research

[a] Though no specific methodological score was proposed, the authors did rank their criteria based on their relative importance. The authors also favor a scoring system that could be used to assign credits/points each time a criterion is present in a study. Several ideas for how to assign scores are proposed.

[b] Development of the methodological scores was based on previous studies (Minnerup et al. 2008; Minnerup et al. 2009). To calculate a quality score, one point is awarded for each quality assessment criterion that is mentioned in a study.

[c] To calculate the quality score, points were awarded if the assessment criteria were satisfied in the article. The scores given for each question were added to give an overall score which was expressed as a percentage of the total possible score. Data were classified as unacceptable (≤ 50%), acceptable (51-79%), or high (≥ 80%).

[d] To calculate the methodological score, one point is given for each criterion if mentioned in the article.

[e] To calculate the methodological score, one point is given for each criterion if mentioned in the article. Studies containing total quality scores less than 5 are considered to be of "poor methodological quality." Studies scoring 5 or 6 points are considered to have "moderate methodological quality" and studies scoring 7 or 8 points are considered to have "good methodological quality."

[f] To calculate the methodological score, one point is given for each criterion if mentioned in the article. Studies scoring less than 4 are considered to be of "poor methodological quality and studies scoring at least 4 points are considered to be of "good methodological quality."

Table 2. Study Design Criteria Aimed at Reducing Bias by Instrument

Instrument Identifier	Random Allocation of Treatment	Allocation Concealment	Blinding	Inclusion Exclusion Criteria Stated	Sample Size Calculation	Compliance with animal welfare requirements	Conflict of Interest Disclosed	Statistical Model Explained	Animals with Co-morbidity	Test Animal Details	Dose Response Model	Every Animal accounted for	Optimal Time Window Used	N (%) of criteria in each instrument (n = 13)
Vesterinen et al. 2011*	Y	Y	Y	Y	Y	N	Y	Y	N	Y	N	Y	N	9 (69%)
Agerstrand et al. 2011*	Y	N	N	N	N	N	N	Y	N	Y	Y	N	Y	5 (38)
National Research Council (US) Institute for Laboratory Animal Research 2011*	Y	N	Y	Y	N	N	N	N	N	Y	N	Y	N	5 (38)
Lamontagne et al. 2010*	Y	Y	Y	N	Y	N	N	N	Y	N	N	N	N	5 (38)
Conrad and Becker 2010*	N	N	N	N	N	N	Y	N	N	N	N	N	N	1 (8)
Vesterinen et al. 2010	Y	N	Y	N	Y	Y	Y	N	N	N	N	N	N	5 (38)
Kilkenny et al. 2010*	Y	N	Y	N	Y	Y	Y	Y	N	Y	N	N	N	7 (54)
Minnerup et al. 2010*	Y	N	Y	N	N	Y	Y	N	Y	Y	N	N	N	6 (46)
Hooijmans et al. 2010*	Y	N	Y	Y	Y	Y	N	Y	N	Y	N	Y	N	8 (62)
van der Worp et al. 2010*	Y	Y	Y	Y	Y	N	N	Y	N	N	N	Y	N	7 (54)
Macleod et al. 2009*	Y	Y	Y	Y	Y	N	Y	N	N	Y	N	Y	N	8 62)
Fisher et al. 2009*	Y	Y	Y	Y	Y	N	Y	Y	Y	N	Y	N	N	9 (69)

Instrument Identifier	Random Allocation of Treatment	Allocation Concealment	Blinding	Inclusion Exclusion Criteria Stated	Sample Size Calculation	Compliance with animal welfare requirements	Conflict of Interest Disclosed	Statistical Model Explained	Animals with Co-morbidity	Test Animal Details	Dose Response Model	Every Animal accounted for	Optimal Time Window Used	N (%) of criteria in each instrument (n = 13)
Rice et al. 2008*	Y	N	Y	N	Y	N	N	N	N	Y	N	Y	N	5 (38)
Sniekers et al. 2008*	N	N	Y	N	Y	N	N	N	N	Y	N	N	Y	4 (31)
Sena et al. 2007*	Y	Y	Y	N	Y	Y	Y	N	Y	N	Y	N	N	8 (62)
Unger 2007	Y	N	Y	N	N	N	N	Y	N	N	N	Y	N	4 (31)
Hobbs et al. 2005*	N	N	N	N	N	N	N	Y	N	Y	Y	N	N	3 (23)
Marshall et al. 2005*	Y	N	Y	N	Y	N	N	N	N	Y	N	Y	N	5 (38)
van der Worp et al.2005*	Y	N	Y	N	Y	N	N	N	Y	N	N	N	N	4 (31)
de Aguilar-Nascimento 2005*	Y	N	Y	N	Y	N	N	N	N	N	N	N	N	3 (23)
Macleod et al. 2004*	Y	N	Y	N	Y	Y	Y	N	N	N	N	N	N	5 (38)
Bebarta et al. 2003	Y	N	Y	N	N	N	N	N	N	N	N	N	N	2 (15)
Verhagen et al. 2003*	N	N	N	N	N	N	N	Y	N	N	Y	N	N	2 (15)
Lucas et al. 2002*	Y	N	Y	N	N	N	N	N	N	N	Y	N	N	3 (23)
Festing and Altman 2002*	Y	N	Y	N	Y	N	N	Y	N	Y	N	N	N	5 (38)
Johnson and Besselsen 2002*	Y	N	N	N	Y	N	N	Y	N	N	N	N	Y	4 (31)
Horn et al. 2001*	Y	N	Y	N	N	N	N	N	N	N	Y	N	N	3 (23)
Durda and Preziosi 2000*	Y	N	N	N	N	N	N	Y	N	Y	Y	N	N	4 (31)

Instrument Identifier	Random Allocation of Treatment	Allocation Concealment	Blinding	Inclusion Exclusion Criteria Stated	Sample Size Calculation	Compliance with animal welfare requirements	Conflict of Interest Disclosed	Statistical Model Explained	Animals with Co-morbidity	Test Animal Details	Dose Response Model	Every Animal accounted for	Optimal Time Window Used	N (%) of criteria in each instrument (n = 13)
Klimisch et al. 1997*	N	N	N	N	N	N	N	N	N	Y	Y	N	N	2 (15)
Hsu 1993*	Y	N	Y	N	Y	N	N	N	N	N	Y	N	N	4 (31)
N (%) of instruments containing each criterion (n =30)	25 (83%)	6 (20)	23 (77)	6 (20)	18 (60)	6 (20)	9 (30)	12 (40)	6 (20)	14 (47)	10 (33)	7 (23)	3 (10)	

Y indicates that the criterion was present. N indicates that the criterion was not present. * indicates that the instrument contained additional criteria (see Supplemental Material, Table S1).

Figure Legend

Figure 1. Flow of Included Studies. N indicates the number of studies.

Figure 1. Flow of Included Studies